SED RDAS Data Preparation Methodology

November 2021

The Survey of Earned Doctorates (SED) Restricted-Use Data Analysis System (RDAS) is a new tool that the National Center for Science and Engineering Statistics (NCSES) has produced to expand public access to SED data. The SED RDAS was designed to produce high-quality estimates while protecting against disclosure of confidential information.

Unlike other SED data products, the RDAS is based on a sample of SED data that have been subject to nonresponse weight adjustment and missing data imputation, as well as statistical disclosure avoidance methods, which modify the data to protect disclosure of confidential information. As a result, the RDAS produces estimates that may differ from published tables; however, the missing data treatment adds value to SED data by allowing users to create tables that provide estimates for the full population using responses from all survey respondents. Further, it makes the number of doctorates in different RDAS tables internally consistent rather than being limited to those who responded to the items represented in each table.

The initial SED RDAS was created using data only for 2018 and 2019. Over time, additional years will be added. SED data are also available in the NCSES Table Builder which contains SED data going back to 1958 but only for a limited number of variables (institution, field of study, race and ethnicity, sex, and citizenship), while the RDAS contains a more comprehensive set of variables (50 data items). The Table Builder is based on the full SED data set, with no imputation, weighting, or alteration for disclosure avoidance. Confidentiality is protected via limitations on the variables available and restrictions on which variables can be tabulated together. Complete SED microdata files, which are not subjected to missing data or disclosure avoidance treatment, are only available under a restricted-use data license.

RDAS Restrictions for Confidentiality Protection

SED RDAS users may notice some limitations on functionality, such as limits on the number of variables that may be used together and certain combinations of variables that are not available. These restrictions exist to protect the confidentiality of individual respondents. By applying these restrictions in tandem with statistical disclosure avoidance methods, the SED RDAS is able to provide information on the population of doctorates without compromising individual records. The key restrictions on functionality are the following:

- Limitation on table dimension. The number of variables that can be crossed (including row, column, and filter) is limited to three. This also limits the proportion of tables subject to suppression.
- Suppression of tables with small cell counts. About 0.5%-1% of two-way tables are suppressed. Excluding tables containing detailed field of study, about 10-11% of three-way tables are suppressed. Three-way tables containing detailed field of study have a high rate of suppression. Other variables with high rates of suppression, particularly in three-way tables, are HBCU status of baccalaureate institution and broad field of study.

Users may be used to seeing individual table cells suppressed for confidentiality; however, complementary cell suppression across the numerous linked tables in the RDAS is not feasible. Therefore, whole tables must be suppressed, even if only part of the table is selected for viewing. Efforts were made to limit the impact of the suppression rule; however, users requiring detailed field tables may find this limiting and should seek other means of access to SED data.

RDAS Missing Data Treatment

There are two types of missing data in SED: unit nonresponse, and item nonresponse. About 8% of doctorates do not participate in SED and are considered unit nonrespondents. Item nonresponse occurs when doctorates who participate in SED do not answer one or more survey questions. In some cases, item nonresponse occurs because respondents were not asked certain questions. This can occur if an SED participant responded by means of an abbreviated telephone interview. Unit nonresponse is addressed via nonresponse weight adjustment, while item nonresponse is addressed by imputation.

Nonresponse weight adjustment. Some information is known for most or all doctorates, regardless of SED respondent status, including age, race, ethnicity, citizenship, sex, field of study, and institution characteristics (Census region, Carnegie classification, historically Black college or university (HBCU) status, public or private control). This information is used to create a nonresponse weight adjustment that allows population inferences to be made from unit respondents using the SUDAAN procedure WTADJUST.[1] The nonresponse adjustment serves to redistribute the weight of nonrespondents among respondents that have similar characteristics. Since the SED is a census without sampling weights, the weights prior to adjustment were equal to 1. Due to the high response rate, the increase in variance due to this adjustment, or unequal weighting effect (UWE), is quite small (UWE = 1.00).

Missing data imputation. For the SED data in the RDAS, missing values for SED unit respondents are imputed using sequential regression with classification and regression tree (CART) models, implemented using the R package mice.[2,3] This is a flexible approach that can automatically detect and model important relationships in the observed data. Logical relationships, or skips, are also preserved in the imputed data. All variables are imputed in one batch, with all RDAS variables as candidate predictors for each imputation model. The imputation rate for most variables is less than 3%.

RDAS Statistical Disclosure Limitation

Several statistical disclosure methods are employed to protect data in the RDAS against disclosure of confidential information. These include coarsening and variable selection, sampling, and disclosure imputation.

Coarsening and variable selection. Variables are coarsened by combining small categories to create larger ones or by creating categorical variables out of continuous variables. Smaller categories present a disclosure risk because they can be used to identify unique individuals, while coarsening makes individuals appear more similar to each other. For continuous variables, outliers present a disclosure risk because they can be used to identify unique individuals. RDAS variable selection and coarsening were determined based on NCSES knowledge of user interest and examination of frequency tables. Coarsening was applied with a goal of creating analytically useful broad categories for which a reasonable number of tables could be produced without RDAS suppression rules preventing too many

tables from being reported. Variables with high suppression rates after coarsening were removed from the variable list.

Sampling. Sampling provides protection against disclosure of the SED data by creating uncertainty about membership inclusion in the data set. Because the SED is a census, this is a change in how the SED data are delivered and analyzed; however, by using a complex sample design and having access to a near-complete population data set, the RDAS is able to use a sample size larger than typical for sample surveys—and thus the RDAS can be considered to provide high-quality SED survey estimates. Standard error estimates and confidence intervals that account for sampling errors are not displayed by default, but are available as an option. These are computed inside the RDAS using bootstrap replicate weights.

The sample design was modeled on the Survey of Doctorate Recipients (SDR), which uses SED as a sampling frame. The design is a stratified systematic random sample using serpentine sort with random start. Explicit and implicit stratification variables were chosen to ensure that the sample was representative on key variables. An optimization procedure was used to allocate the sample across explicit strata while meeting as many precision requirements as possible. The final analysis weight used in the SED RDAS that accounts for the nonresponse adjustment and sampling has a UWE of 1.02.

Bootstrap replicate weights were created to facilitate variance estimation in the RDAS. The type of replicate weight created is Rao and Wu's *n-1* bootstrap.[4] Two hundred replicates were created using the R package survey.[5] The Rao-Wu bootstrap is appropriate for both with replacement and without replacement designs; yields sensible variance estimates for a variety of estimators, including percentiles; and can produce adequate estimates when sample sizes are small.[6]

Perturbation. While sampling creates uncertainty about membership inclusion in the data set, it does not eliminate the risk of disclosure. Best practices for protecting sample surveys against re-identification of confidential information include using a perturbative disclosure protection method that has a random component. Perturbative methods add uncertainty to any attempted re-identification, ideally while preserving analytical properties of the data.

The perturbation method used was disclosure imputation, which is similar to the method used for missing data imputation.[7] That is, a portion of values were replaced with values obtained via imputation models. Imputation was conducted using CART models and the R package synthpop,[8] which is specifically designed to do imputation, or synthetic data, for disclosure protection. The variables and records selected for perturbation are confidential. All RDAS variables were supplied to the CART algorithm as candidate predictors for each imputation model, so that relevant predictors and interaction terms would be included in the models, thus preserving analytic properties of the data. Quality control and data validation procedures were used to verify that the perturbation did not create undue noise by confirming the validity of confidence intervals and verifying that coefficients of variation (CV) were less than 0.5; less than 0.1% of estimates have CV greater than 0.5, and only in cells with fewer than 25 doctorates.

[1] RTI International. 2012. *SUDAAN User's Manual, Release 11.0*. Research Triangle Park, NC: RTI International.
[2] Van Buuren S, Groothuis-Oudshoorn K. 2011. Multivariate imputation by chained equations in R. *Journal of Statistical Software* 45(3):1–67.
[3] Burgette LF, Reiter JP. 2010. Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology* 172(9):1070–6.
[4] Rao JNK, Wu CFJ. 1993. Bootstrap inference for sample surveys. *Proceedings of the Section on Survey Research Methodology*, 866–71.
[5] Lumley T. 2004. Analysis of complex survey samples. *Journal of Statistical Software* 9(1):1–19.
[6] Girard C. 2009. The Rao-Wu rescaling bootstrap: From theory to practice. *Federal Committee on Statistical Methodology Proceedings.*
[7] Kinney SK, Looby CB, Yu F. 2020 Advantages of imputation vs. data swapping for statistical disclosure control. *In* Domingo-Ferrer J, Muralidhar K, editors, *Privacy in Statistical Databases. PSD 2020. Lecture Notes in Computer Science,* Vol. 12276, pp. 281–96. Cham, Switzerland: Springer.
[8] Nowok B, Raab GM, Dibben C. 2016. synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical software* 74(11):1–26.